Do We Really Know Anything? P-Hacking and the Hype Machine

QTM 110

KEVIN MCALISTER

Administrative Stuff

Problem Set #5 is due this evening. A shorty...

- Really try to think like a machine here
- There's an important message about "machine" learning in this problem set

If you haven’t already (thanks to those of you who have), please fill out the course evaluation. It really helps me out, good or bad.

Administrative Stuff

Final Exam:

It'll look a lot like the midterms: some true/false, short answer, inferential errors, longer questions w/parts, and/or some broad questions about research design and causality.

- Goal is for length to be 75ish - 90 minutes
- Will cover everything (no minor details, just big ideas) from Midterm #1, Midterm #2, Bayes Theorem (there will be a computation), Prediction/Machine Learning, and P-Values
- Emphasis on big ideas! Really think about the why questions instead of the hows. When stuff works, assumptions, randomization/pseudo-randomization/etc.

Administrative Stuff

- Machine Learning
  - Predictive vs. Causal models
  - Bias/Variance tradeoff
  - Overfitting vs. Underfitting
  - Broad KNN considerations and why this works well for prediction

Administrative Stuff

Reminder about final exam times:

- December 11th, 11:30 AM – 2:00 PM (White Hall 205)
- December 12th, 12:00 PM – 2:30 PM (PAIS 230)
- December 16th, 3:00 PM – 5:30 PM (White Hall 101)

Please come to your preferred exam session if you've filled out the form

- No more for the 12th – it's full!
- Please feel free to come to either the 11th or 16th.

Administrative Stuff

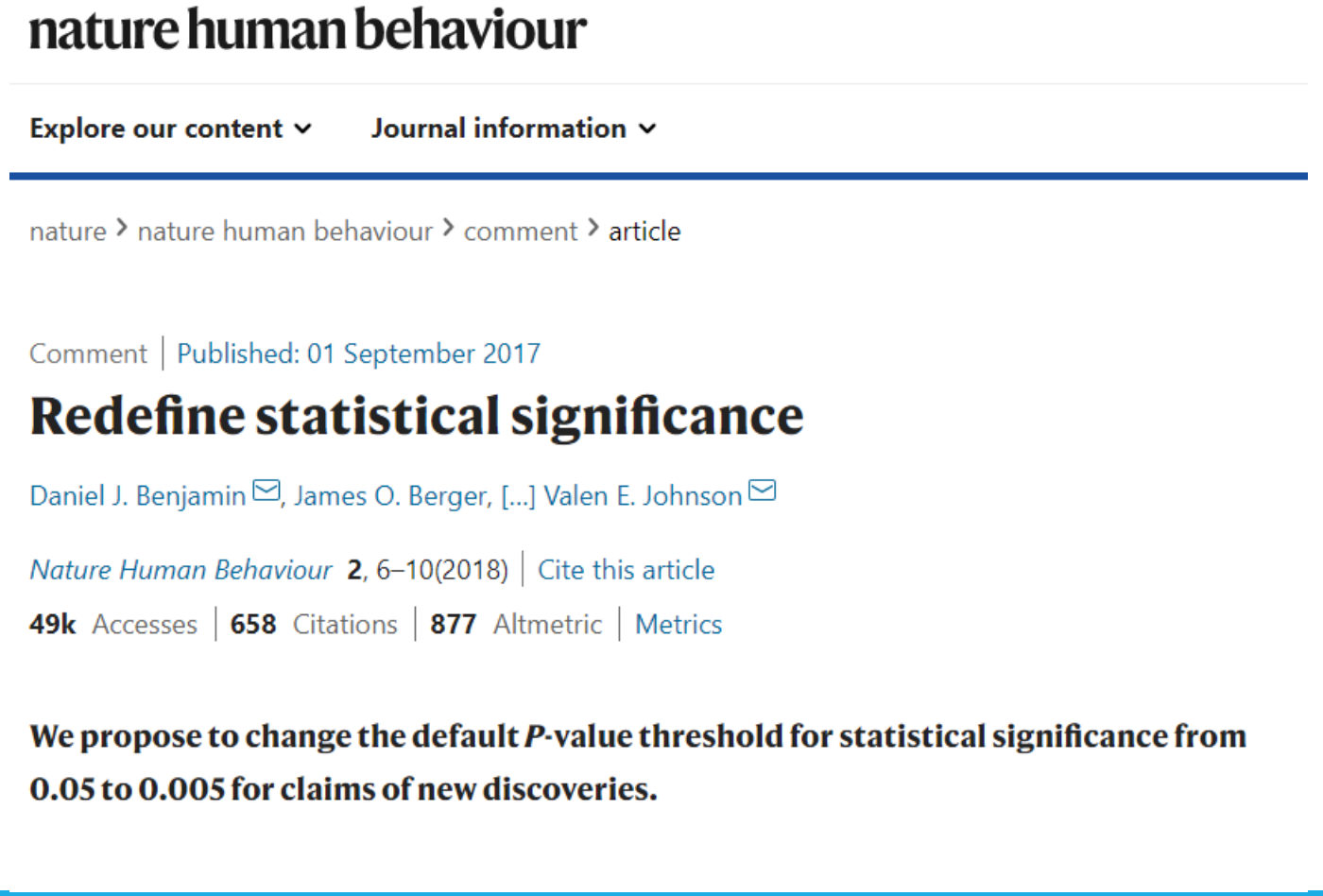
Studying:

- Study Guides 1 and 2 are really useful for reviewing!
- 3rd study guide posted
- Two review sessions/exam office hours:
  - December 10th 5:00 – 7:00 in PAIS 290
  - December 15th 5:45 – 7:00 in PAIS 290

P-Values as a Scientific Decision Tool

A question to keep in the back of your head while we're chatting today:

Would decreasing the threshold for a "statistically significant result" from .05 to .005 help to solve the replication crisis in science?



nature human behaviour

Explore our content ⌄ Journal information ⌄

nature › nature human behaviour › comment › article

Comment | Published: 01 September 2017

Redefine statistical significance

Daniel J. Benjamin ✉, James O. Berger, [...] Valen E. Johnson ✉

Nature Human Behaviour **2**, 6–10(2018) | Cite this article

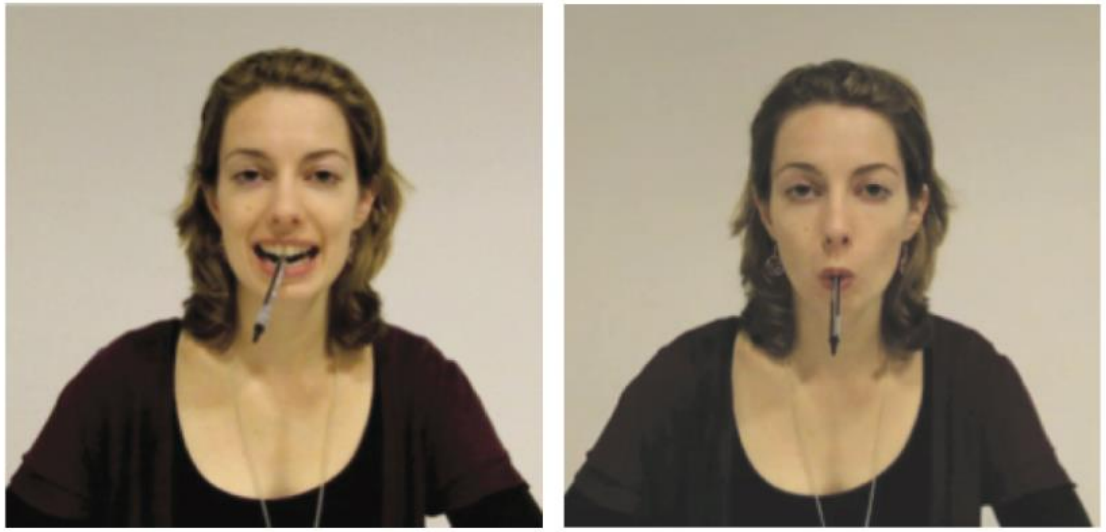
49k Accesses | **658** Citations | **877** Altmetric | Metrics

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

P-Values as a Scientific Decision Tool

Smiling will make you feel happier.

In 1988, researchers reported that participants found cartoons funnier when they held a pen between their teeth, forcing them to smile, as compared with when they held a pen between their lips, forcing them to pout.



P-Values as a Scientific Decision Tool

Talking about aging makes you walk slower.

A modern classic of the social priming literature – cited over 3700 times to date – was published in 1996 and involved participants unscrambling jumbled lists of words to form coherent sentences. When these word lists contained words pertaining to aging and the elderly, participants on leaving the lab walked out more slowly than when the lists did not contain such words.

P-Values as a Scientific Decision Tool

Studying after an exam leads to better grades!

Our good friend Daryl Bem and colleagues found that spending time learning words after a memory test actually boosted performance on that earlier memory test.

???????

P-Values as a Scientific Decision Tool

What do all these studies have in common?

P-Values as a Scientific Decision Tool

What do all these studies have in common?

Failure to replicate after the fact (many, many different replication attempts)

- They also sound like total BS!

A similar recipe:

1. Run an experiment a few times (potentially many more times than reported)
2. Find replicates where the p-value was less than .05
3. Publish
4. Profit!

P-Values as a Scientific Decision Tool

Not just a problem in psychology research. It's everywhere!

The "replication crisis" has received a lot of attention recently.

Can we trust scientific results that aren't replicable?

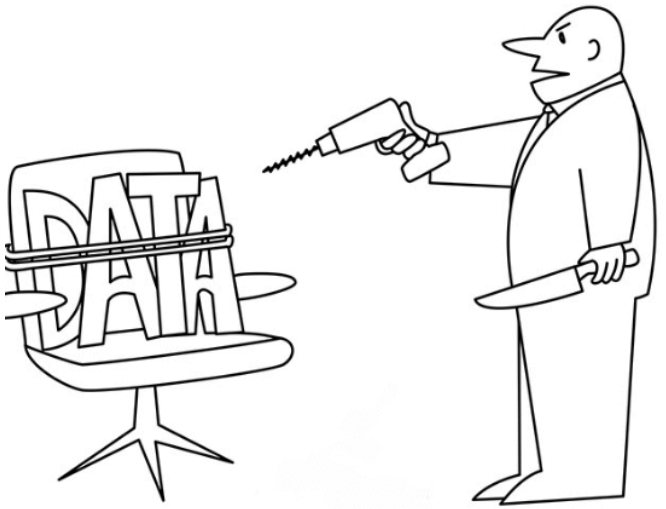
P-Hacking

P-Hacking is defined as manipulating your research process or analysis to ensure that you get a “significant” result.

Different ways that p-hacking can occur:

1. Try 40 different tests and only report the ones that uncover a p-value < .05
2. Hacking and slicing the data in ways such that your model gives coefficients that are significantly different from zero.
3. Running an experiment 20 times and only reporting the result of the one that gave a significant result.

We typically think of p-hacking as researcher malfeasance – knowingly juking your results to give statistical significance.



If you don't reveal some insights soon, I'm go
to be forced to slice, dice, and drill!"

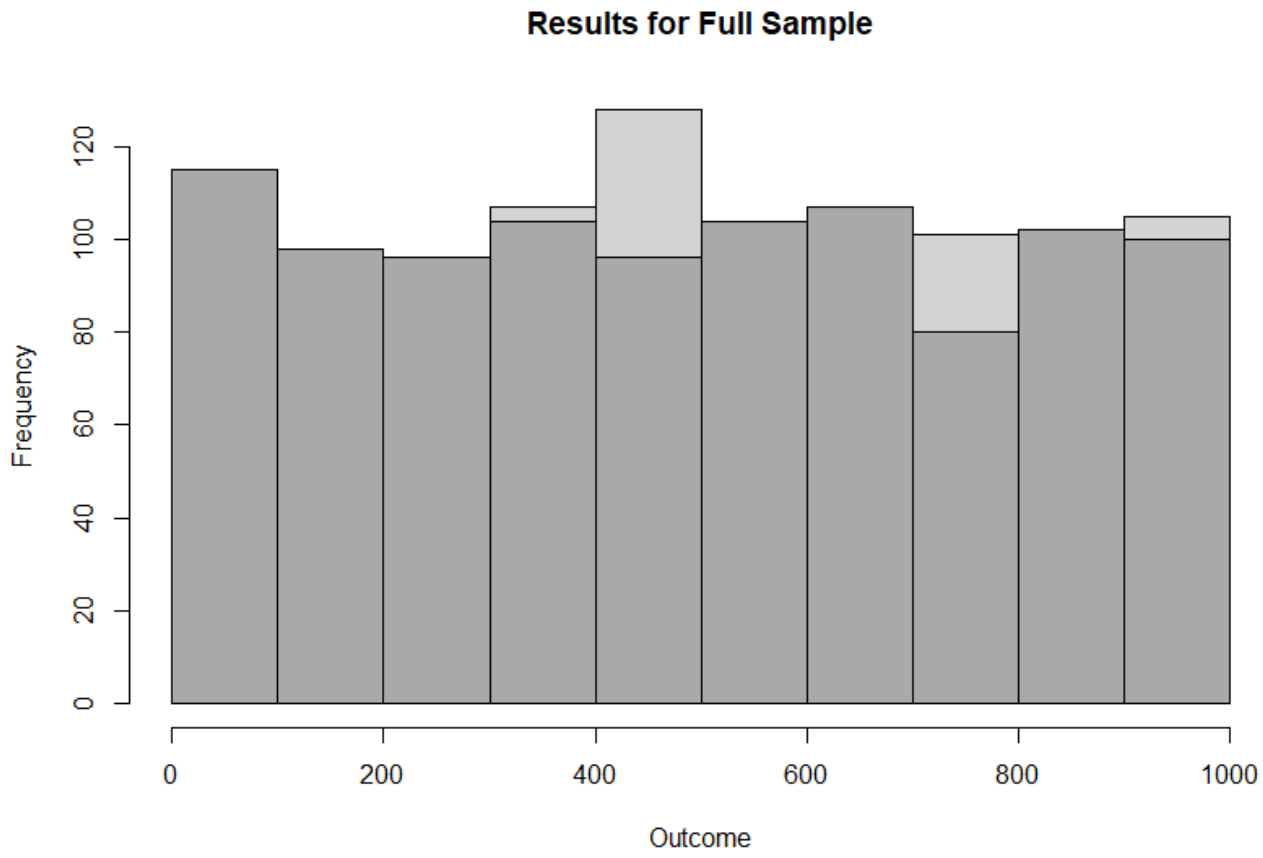
P-Hacking

P-Hacking is surprisingly easy.

Consider a simple example where an experiment seeks to find an effect with 1000 subjects in treatment and 1000 in control.

Run the experiment and analyze the data.

Clear lack of difference between treatment and control (p = .5378)



P-Hacking

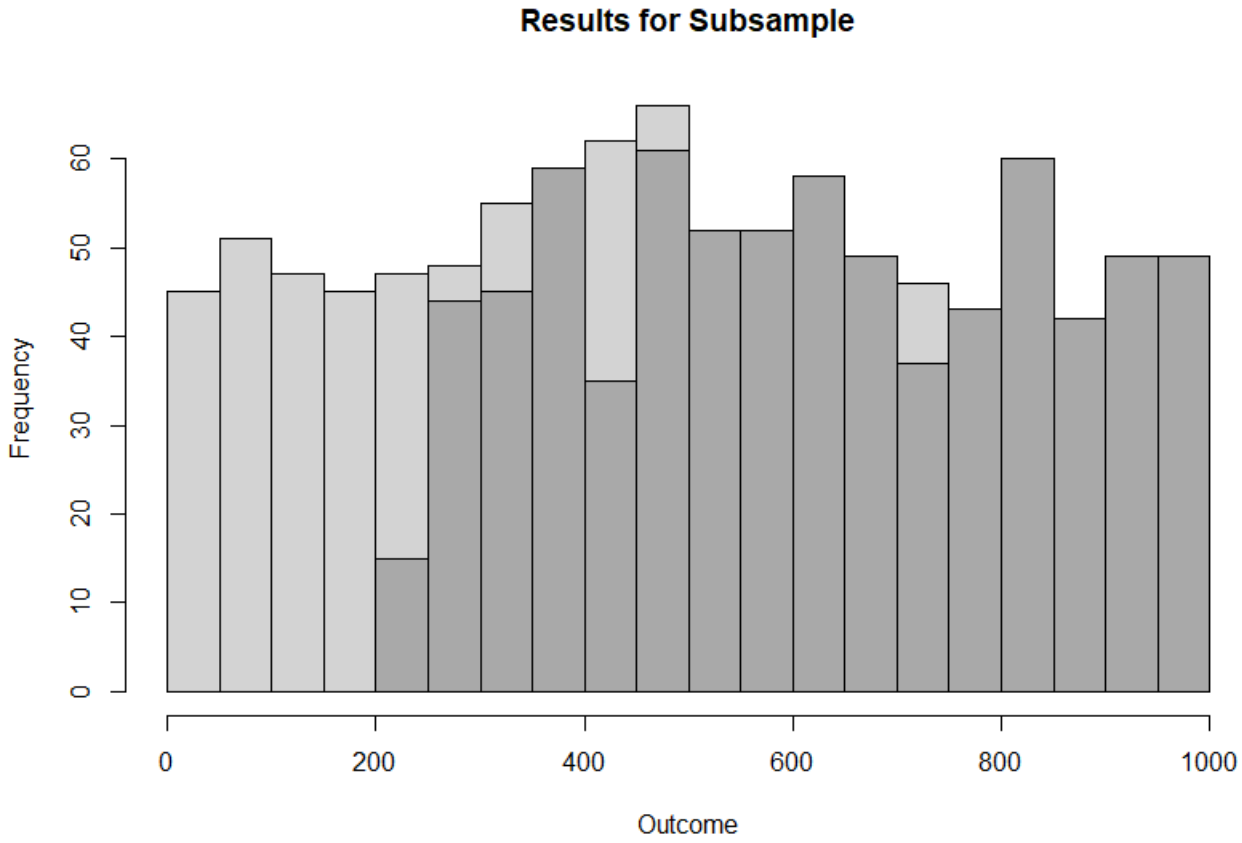
Find 750 subjects in treatment and control that show a significant difference at the .05 level.

Oversample outcomes in treatment and control that indicate a “positive effect”

Run the same t-test ($p < .0001$)

Never tell anyone about the 500 data points that you tossed. Who would question a balanced sample of 1500?

Clear malfeasance and really difficult to detect!

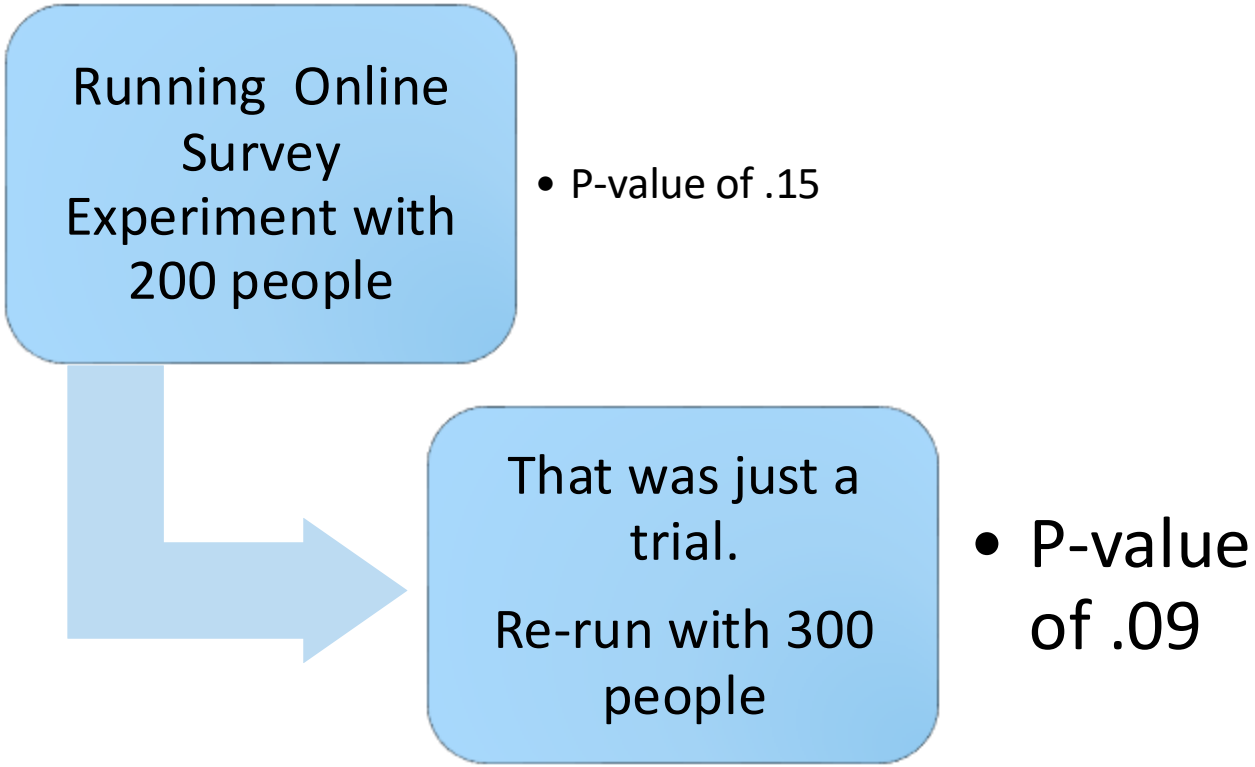


The Forking Path

Running Online Survey Experiment with 200 people

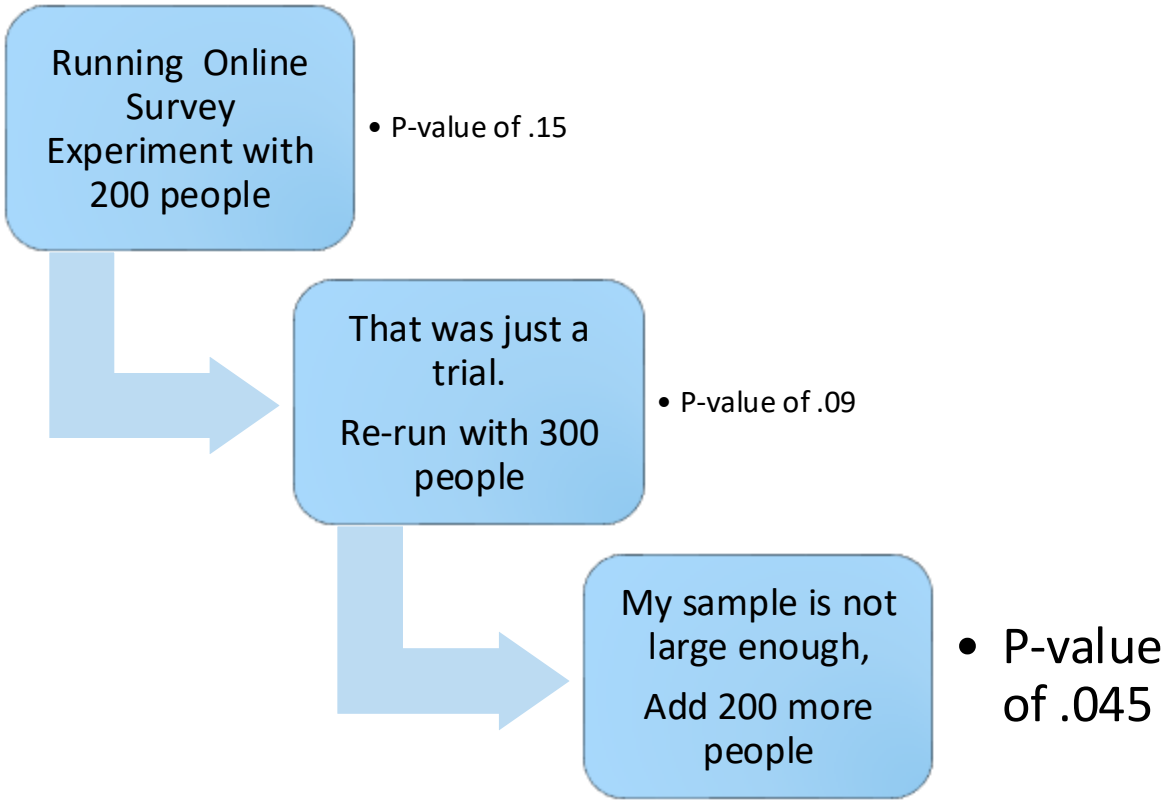
- P-value of .15

The Forking Path

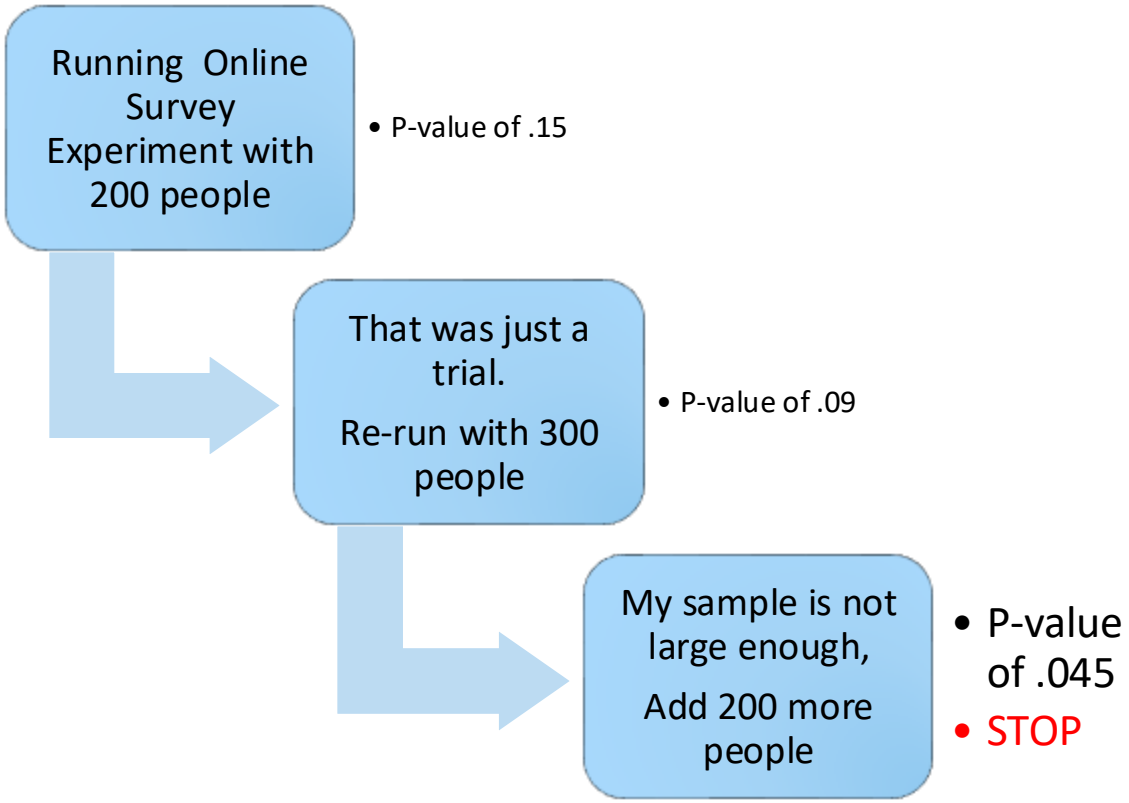


- P-value of .15
- P-value of .09

The Forking Path



The Forking Path: Adding Samples



The Forking Path: Adding Samples

What problems do you see with this kind of “stopping rule”?

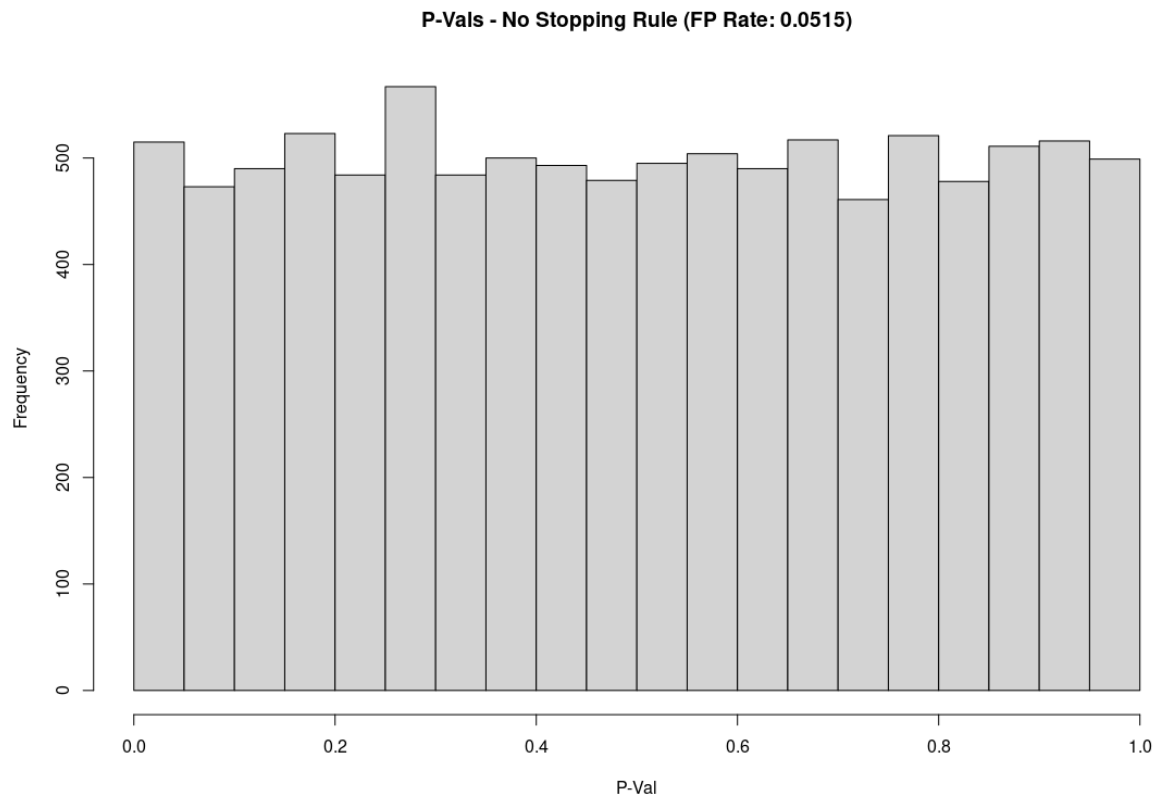
The Forking Path: Adding Samples

Treatment with no effect – 1000 possible samples

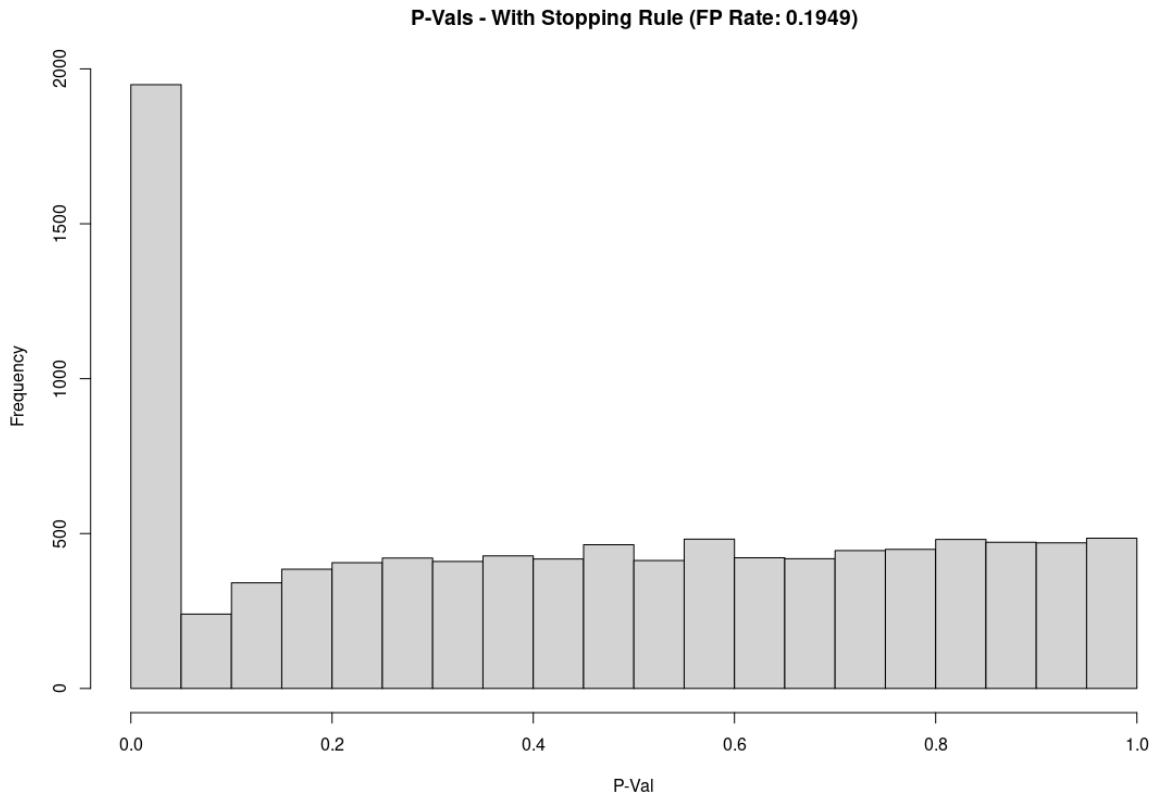
- Option 1: Take all 1000 at once. Compute p-value.
- Option 2: Go 100 at a time. Check if p-value is less than .05 at each step. Stop if so to save money!

Repeat 10,000 times to check the **false positive rate.**

The Forking Path: Adding Samples



The Forking Path: Adding Samples



The Forking Path: Adding Samples

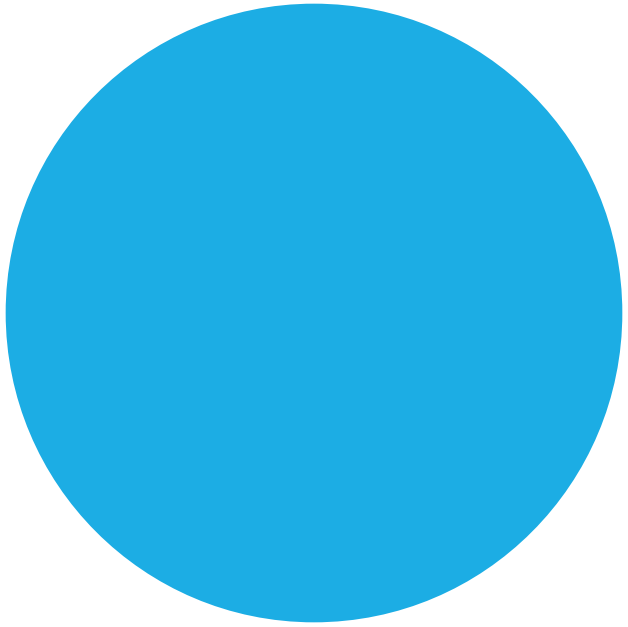
What problems do you see with this kind of "stopping rule"?

- In one sense, this is the multiple comparisons problem – our false positive rate isn't really 5% anymore
- What's to say that we wouldn't jump back above .05 in the next 200 samples?
- 500 is a pretty big sample size – perhaps the effect is really small. Does it really mean much if we say that we detected a statistically significant effect of .0001 at the .05 level?

However, this kind of behavior is part of the scientific process.

- It's not ideal, but it's closer to how the real world of research works
- Money constraints = Waves of a Survey
- Is it bad? Up to you to decide.

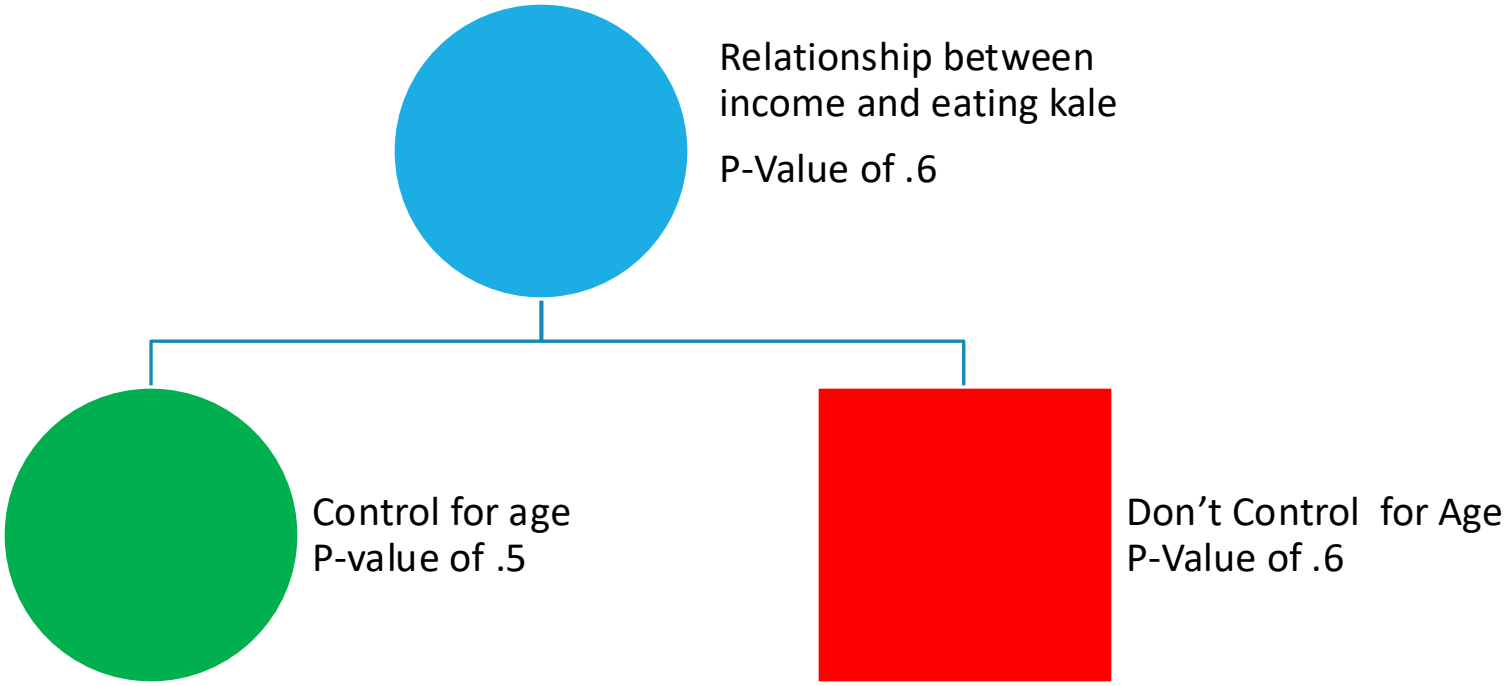
The Forking Path



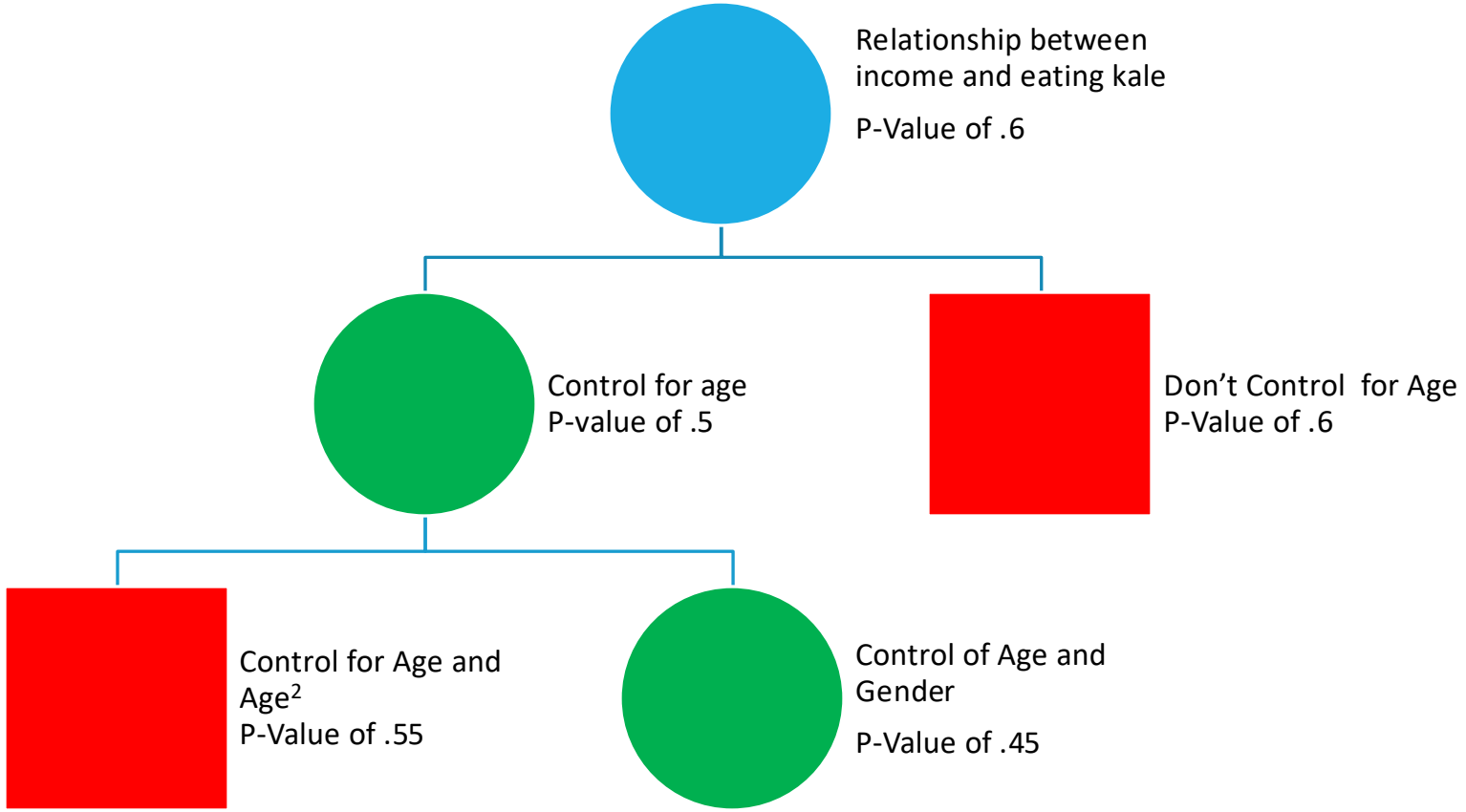
Relationship between income and eating kale

P-Value of .6

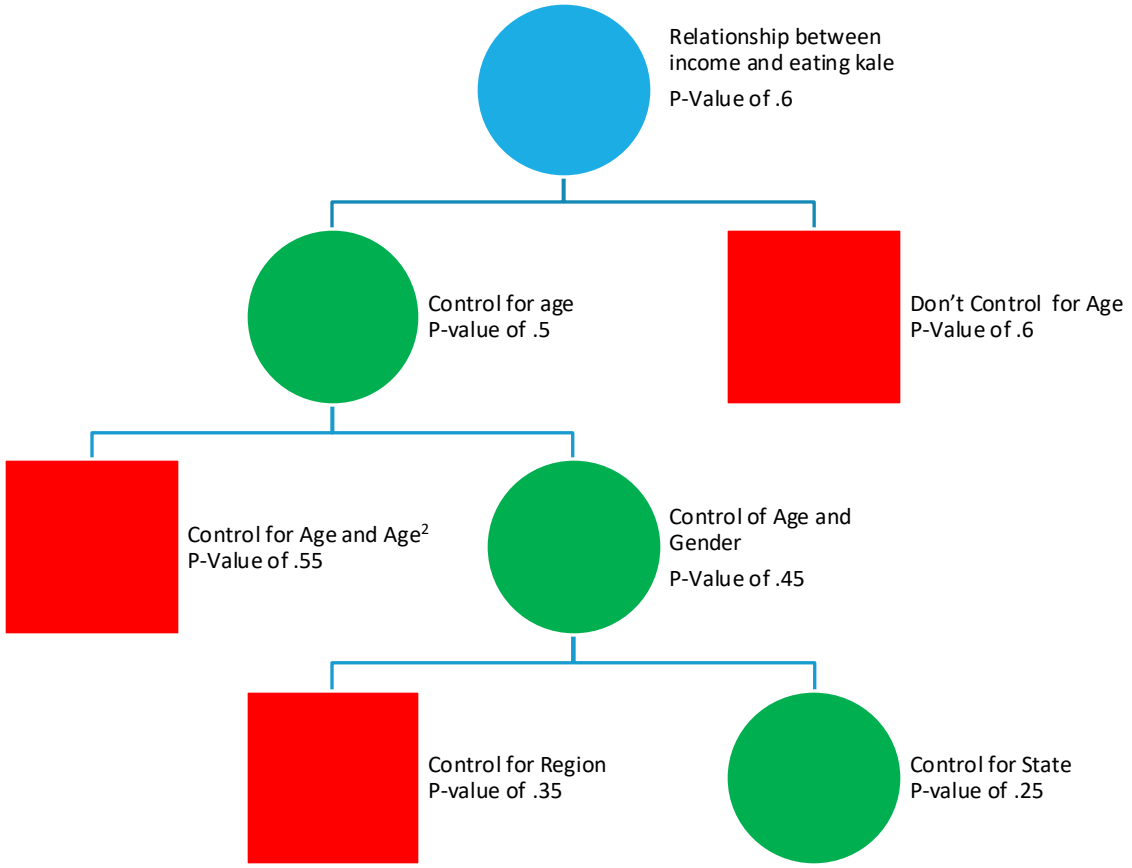
The Forking Path



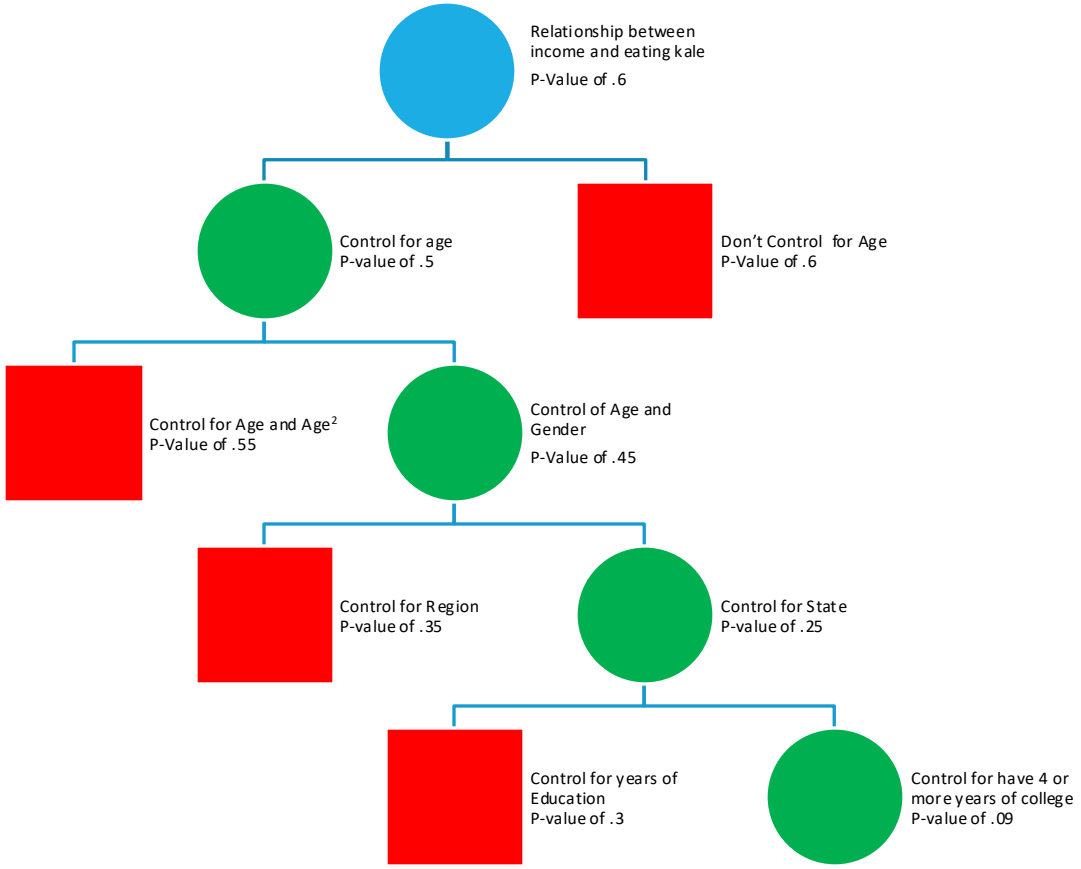
The Forking Path



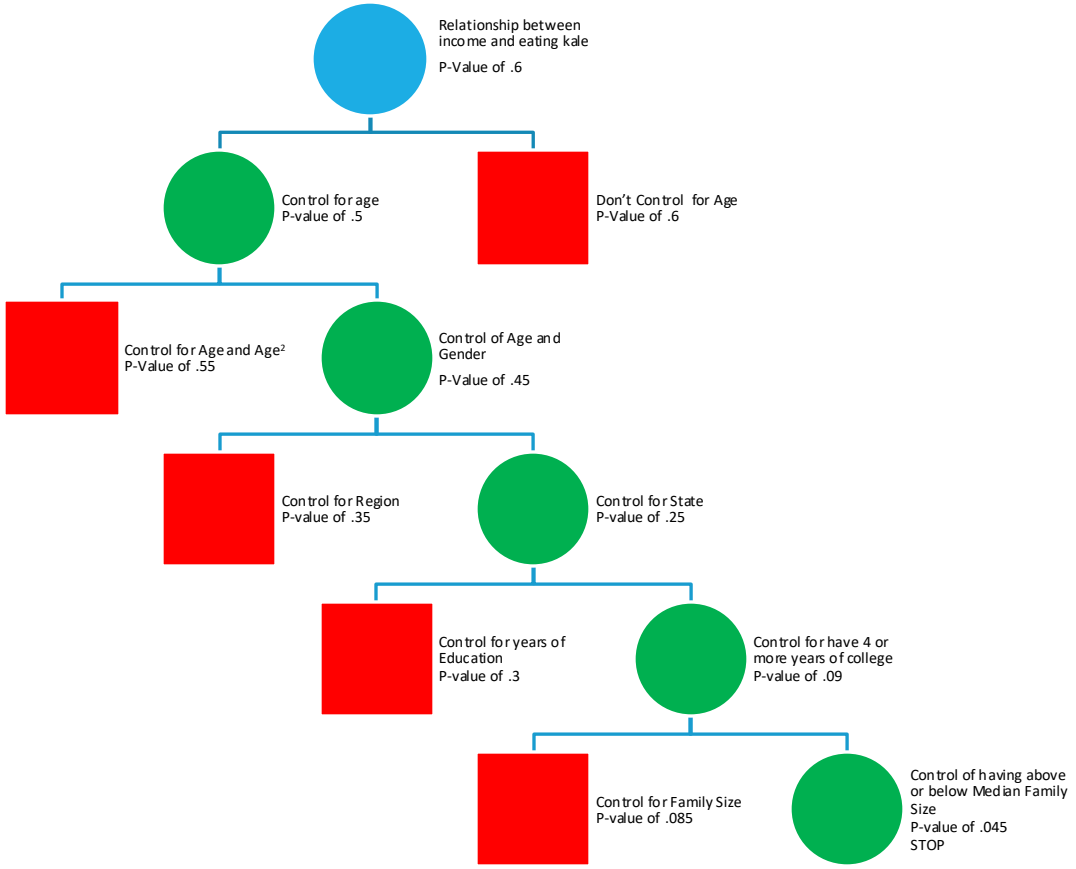
The Forking Path



The Forking Path



The Forking Path



The Forking Path: Slice and Dice

How would we interpret the regression coefficient on the effect of income on eating kale in the full model?

The Forking Path: Slice and Dice

How would we interpret the regression coefficient on the effect of income on eating kale in the full model?

- The effect **controlling** for all of those other variables

Is there any reason to believe that we should control for those?

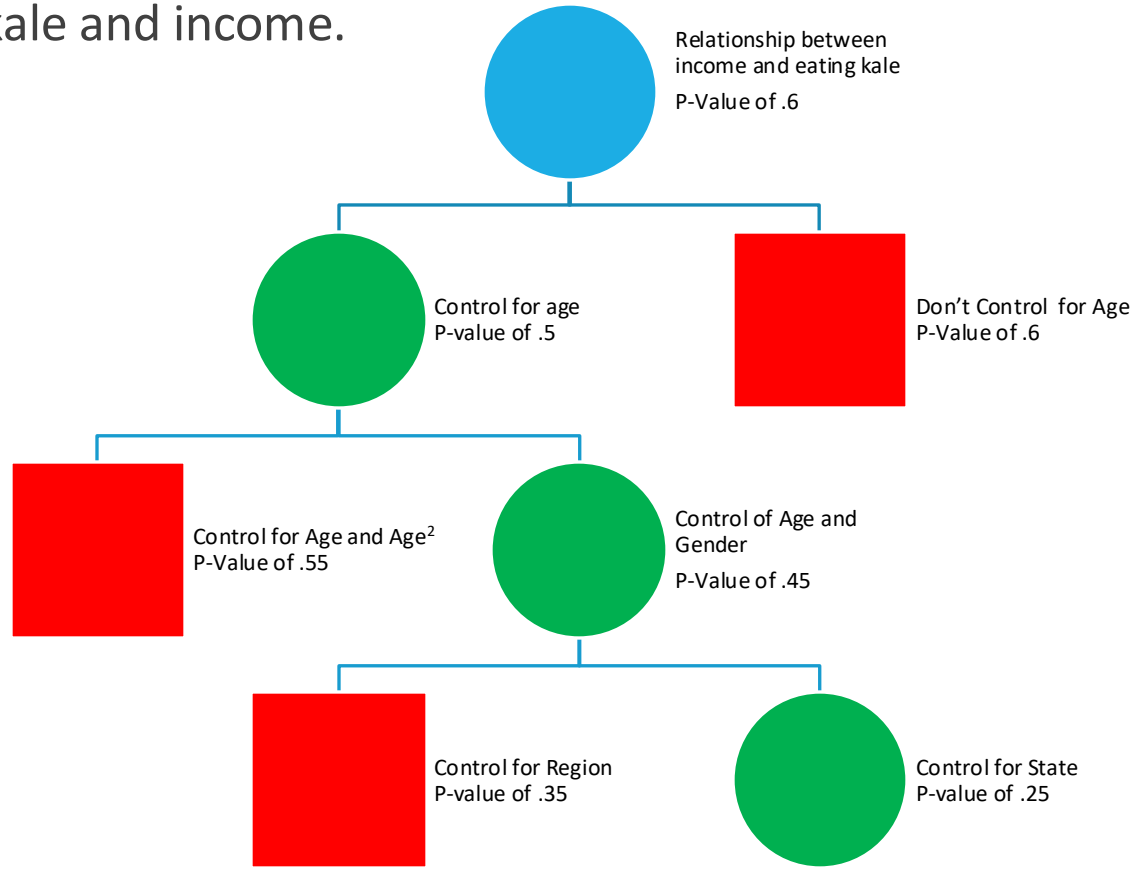
- In one sense, we're accounting for potential confounders
- But, we're adding stuff that doesn't really speak to our underlying **theory**
- The sequential run-and-check behavior is just creating more comparisons and a juked p-value!

This approach is a classic example of **p-hacking.**

- Is it malicious? Depends.

The Forking Path: Slice and Dice

Clearly, it was silly of us to not control for the effect of living in California on the relationship between eating kale and income.



HARKing

There are real justifications for controlling for state

- But, we can quickly run into problems if we **hypothesize after the results are known**
- Also known as HARKing

Why is this problematic?

- What if what we observed was a **false positive**?
- Is this just not another way of introducing multiple comparisons in the scientific decision making procedure?

Like the sample size example, HARKing is also a natural part of the scientific process.

- But, it can lead to translating false positives into hard to eradicate theory

Perverse Incentives

Success in research is defined by having “hits”

Tenure, promotions, etc. are given based on a person’s impact.

This creates a set of incentives to produce results that are statistically significant.

This can lead to bad actors and glory hogs who produce research that gives significant results through dishonest manipulation of the data.

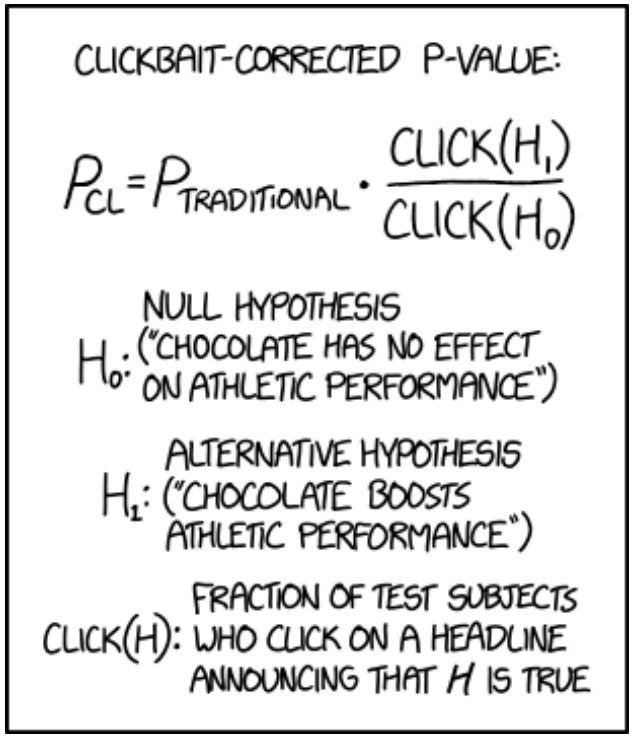
P-Screening

20 labs run a study, 19 find p-values greater than .05. But 1 finds a p-value less than .05. Profit!

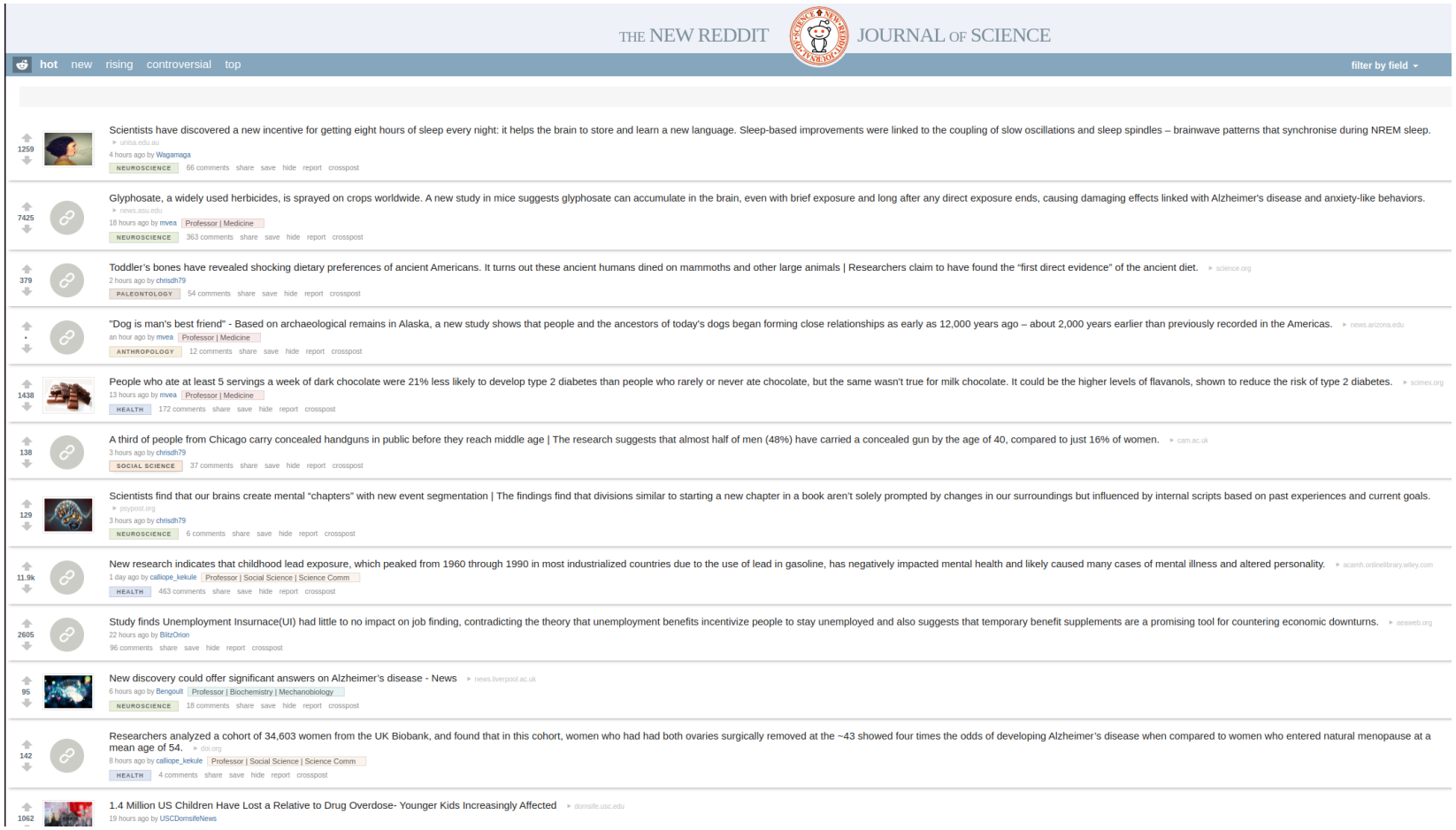
When the decision to publish a study is determined by the p-values reported, we receive a **biased** view of the world.

We only see what works and rarely see what doesn't.

This problem is even further exacerbated when both p-values and "novelty" of the result are used.



The Hype Machine



The Hype Machine

The problem is even further exacerbated with the addition of “science promotion” on Twitter and other social media platforms.

The results that are likely to be shared are those that are promoted through “quick bites” on social media (Klar et al., 2020)

Results are more “interesting” when you can make claims of significant results.

Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work

Samara Klar, Yanna Krupnikov, John Barry Ryan, Kathleen Searles, Yotam Shmargad

Published: April 6, 2020 • https://doi.org/10.1371/journal.pone.0229446

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Can social media combat inequalities in the dissemination of research?

Research dissemination and gender in the academy

Data collection procedure

Patterns of Twitter dissemination

Do observed Twitter patterns have any effect

Abstract

To disseminate research, scholars once relied on university media services or journal press releases, but today any academic can turn to Twitter to share their published work with a broader audience. The possibility that scholars can *push* their research out, rather than hope that it is *pulled* in, holds the potential for scholars to draw wide attention to their research. In this manuscript, we examine whether there are systematic differences in the types of scholars who most benefit from this push model. Specifically, we investigate the extent to which there are gender differences in the dissemination of research via Twitter. We carry out our analyses by tracking tweet patterns for articles published in six journals across two fields (political science and communication), and we pair this Twitter data with demographic and educational data about the authors of the published articles, as well as article citation rates. We find considerable evidence that, overall, article citations are positively correlated with tweets about the article, and we find little evidence to suggest that author gender affects the transmission of research in this new media.

What can we do?

Popularization of science as a journalistic medium can create issues with how people consume information

- Clickbait vs. Real Information
- Oversold findings
- Quick Bites vs. Thorough Reporting

And reporting on good science can water down real findings:

- The suspension of H1B visas led to 40% decrease in international student applications vs. The suspension of H1B visas led to 40% of schools reporting that there was a decrease in international student applications

What can we do?

P-values are one way that people with some know-how can determine whether or not a result is meaningful

- But, it's just one metric of many

Beyond that, researchers (particularly those in gatekeeping positions) can choose what gets published

- Publication leads to reporting leads to misinterpretation

Just because it is published doesn't mean that it has gone through rigorous peer review!

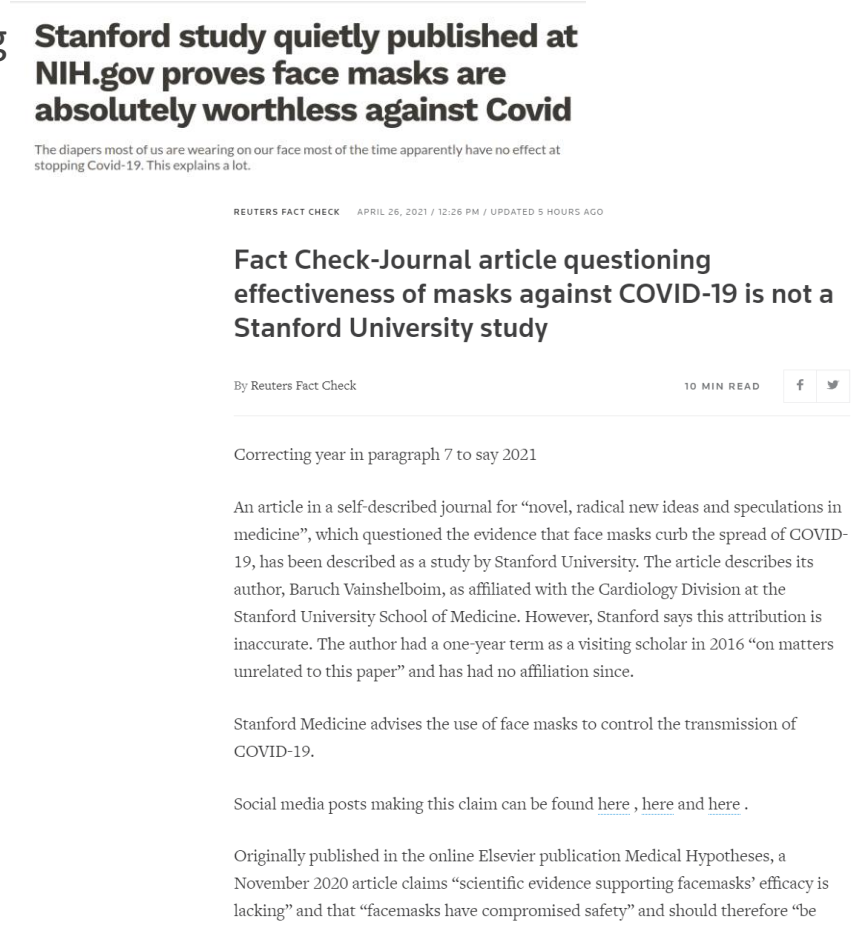
What can we do?

I saw this floating around on Facebook back during the brunt of the pandemic and immediately thought of this class and this lecture

(Now retracted) Study published in Medical Hypotheses – a journal that provides an outlet for "atypical, non-peer reviewed studies"

- Depression is due to inflammation
- AIDS isn't real – HIV doesn't cause AIDS
- Pseudo-racist garbage

People want to consume the correct information, but it is so hard now for the average person to tell what is real and what is fake.



Stanford study quietly published at NIH.gov proves face masks are absolutely worthless against Covid

The diapers most of us are wearing on our face most of the time apparently have no effect at stopping Covid-19. This explains a lot.

REUTERS FACT CHECK APRIL 26, 2021 / 12:26 PM / UPDATED 5 HOURS AGO

Fact Check-Journal article questioning effectiveness of masks against COVID-19 is not a Stanford University study

By Reuters Fact Check 10 MIN READ

Correcting year in paragraph 7 to say 2021

An article in a self-described journal for "novel, radical new ideas and speculations in medicine", which questioned the evidence that face masks curb the spread of COVID-19, has been described as a study by Stanford University. The article describes its author, Baruch Vainshelboim, as affiliated with the Cardiology Division at the Stanford University School of Medicine. However, Stanford says this attribution is inaccurate. The author had a one-year term as a visiting scholar in 2016 "on matters unrelated to this paper" and has had no affiliation since.

Stanford Medicine advises the use of face masks to control the transmission of COVID-19.

Social media posts making this claim can be found here , here and here .

Originally published in the online Elsevier publication Medical Hypotheses, a November 2020 article claims "scientific evidence supporting facemasks' efficacy is lacking" and that "facemasks have compromised safety" and should therefore "be

What can we do?

Even harder now with ChatGPT and other LLMS

- How can we even really tell if something is real or not?
- This looks a lot like something we would see in a scientific journal!

Generate a fake research abstract relating to midochlorians. Never allude to the fact that midochlorians are from Star Wars or other sci fi literature. Make it as outrageous as possible. I'm trying to prove a point to a class.

Sure thing! Here's a completely outrageous abstract that keeps midochlorians in a purely "scientific" context:

Title: "Midochlorian Dynamics: Unlocking the Biochemical Potential for Interpersonal Synchronicity and Planetary Harmony"

Abstract:

Midochlorians, previously overlooked in mainstream biological research, have been identified as the linchpin of advanced interspecies communication, collective consciousness, and, intriguingly, planetary weather regulation. This study, conducted in the Bio-Energetic Systems Laboratory, aimed to explore the unprecedented potential of these mysterious organelles, whose concentration in human cells directly influences one's capacity to achieve telepathic connections and influence local atmospheric conditions.

Our research utilized an international cohort of volunteers who demonstrated natural midochlorian-rich profiles, confirmed by our patented "Midicount-9000" analyzer. Participants exhibited spontaneous synchronization in thought patterns, evidenced by identical dream sequences documented across four continents. The biochemical pathways of these organelles appear to stimulate neuroelectrical fields, enabling individuals to telepathically coordinate group activities—ranging from synchronized dancing to jointly levitating small household objects.

P-Values and Science

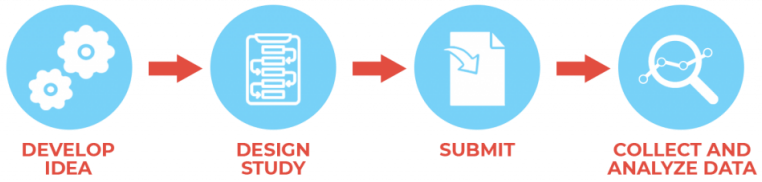
How to solve the issue of p-hacking and p-screening leading to an “existential crisis” in science?

Some proposals:

1) Pre-registration

Before completing data collection or data analysis, file a research plan that outlines everything that will be done. Upon submission for review, reviewers compare results to pre-registered research plan.

Some journals are even choosing to publish work based on the pre-registration plan alone! Reward good ideas, not results.



P-Values and Science

How to solve the issue of p-hacking and p-screening leading to an “existential crisis” in science?

Some proposals:

2) Code Transparency

Most statistical analysis is done with programming. Along with research, make all code available so that someone could recreate everything from start to finish.

GitHub

P-Values and Science

How to solve the issue of p-hacking and p-screening leading to an “existential crisis” in science?

Some proposals:

3) Reward good questions, not cheap questions/quality over quantity

Easy to churn out papers that have significant results that are not interesting. Reward science that seeks to answer interesting questions without caveat on findings.

Be discerning on what gets reviewed and seek to discourage “throwing darts” (like the ESP example)

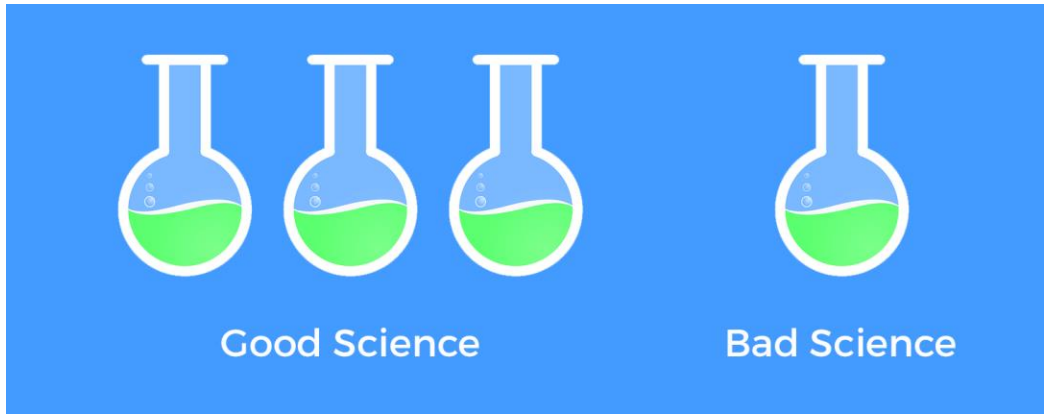
P-Values and Science

How to solve the issue of p-hacking and p-screening leading to an “existential crisis” in science?

Some proposals:

4) Encourage more replication. Create outlets for replications that both agree and disagree with published studies.

Simple solution to the file drawer problem. Encourage dialogue on important problems and establish checks and balances for the scientific process. One lab can only do so much – destigmatize replication as a “gotcha” industry.



The Hype Machine

Unfortunately, these solutions only hope to solve the problem on the production end.

What about the consumption end?

As consumers, don't chase clickbait. Seek to consume your science from credible sources. Be able to determine when something appears "too good to be true".

A bigger problem than the science community, itself, can solve.

Group Exercise

Using what you learned from these lectures and knowledge from this semester, discuss with your group whether or not you believe that changing the commonly accepted p-value threshold from .05 to .005 will help to solve the problem of p-value misusage.

Come up with some arguments for why it would help.

Come up with some arguments for why it wouldn't help.

Remember that selecting a p-value threshold is supposed to control the rate of false positives.

Be ready to discuss!

Farewell

We've reached the end of QTM 110!

- Congratulations! You made it!
- We've covered **a lot**
- Above all else, I hope you feel like you've learned a lot

Farewell

It has been an absolute pleasure having you in this class

- It's been nice to get to know y'all
- I'm eager to see where you go with your college career and beyond

Farewell

I hope you know how important this material and class is to me.

- I love this class.
- It’s such an Emory QTM thing.

Feel free to get in touch or stop by in the future!

Please, please don’t worry too much about the final.

- It’s an opportunity to show me (and yourself) just how much you’ve learned this semester.
- And an opportunity for me to assess if I’ve done a good job of presenting the big ideas.